# Verifying Neural Network Robustness with Dual Perturbations

## Supplementary Material

## 1. Detailed VeriDou Performances

### 1.1. VeriDou Performances on Discrete Kernels

Tab. 1 shows verification performance across different motion blur angles. Due to the limited expressiveness, restricted perturbations fail to capture possbile adversarial examples, thus, achieve significantly higher number of safe (UNSAT) instances (*e.g.*, 156–185 UNSAT instances) compared to dual ones (*e.g.*, 53–126 instances) using $\texttt{VeriDou}_{NS}$. The stability of verification results across different angles suggests that network robustness is not tied to specific geometric properties but rather reflect systematic properties of perturbations.

In contrast, dual perturbations consistently achieve more SAT instances than others (3x) with $\texttt{VeriDou}_{VS}$ showing the largest improvements (8x). Since only correctly classified images are used as input properties, the number of UNSAT instances is proportional to robustness accuracy (*e.g.*, the fraction of UNSAT instances over total instances). Notably, the results expose a fundamental asymmetry that networks appearing to have high robustness accuracy to restricted transformations (*e.g.*, $\texttt{VeriDou}_{\alpha\beta}$ achieves 56.3% robustness accuracy under restricted perturbations at 0°) become vulnerable against dual perturbations (*e.g.*, achieves merely 16.7% robustness accuracy in an identical setup).

### 1.2. VeriDou Performances on Continuous Kernels

Tab. 2 shows results for continuous angle ranges (0→30°, 30→60°, 60→90°). Restricted perturbations claim that networks are highly robust (*e.g.*, 45.0%–63.7% robustness accuracy under $\texttt{VeriDou}_{NS}$) due to their limited space at specific angles, thus, missing potential adversarial examples at intermediate orientations. However, VeriDou reveals substantially more violations (*e.g.*, 135–161 SAT instances), illustrating that counterexamples indeed exist at intermediate angles and surrounding regions that restricted ones overlook entirely. Independent part is particularly effective at further discovering hidden vulnerabilities (*e.g.*, 142→171 under $\texttt{VeriDou}_{\alpha\beta}$) that neither restricted nor universal perturbations could.

### 1.3. VeriDou Performances on Arbitrary Kernels

Tab. 3 presents results using diverse convolutional kernels across three different domains with number of $z_i$ ranging from 20% to 100%. The results show a balanced distribution of UNSAT and SAT instances (*e.g.*, at 20% coverage, UNSAT ranges 101–143 while SAT ranges 81–99). The verification becomes more challenging as kernel parameter coverage increases, thus, timeout instances become increasingly prevalent, *e.g.*, under $\texttt{VeriDou}_{NS}$, timeouts of Domain 3 increase from 75 instances at 20% to 165 instances at 100% coverage.

## 2. Expressivity Theorems

**Lem. 1** (Kernel Space Containment)**.** *Let $\mathcal{K}_{res}$ and $\mathcal{K}_{uni}$ be sets of all kernels expressible by restricted and universal perturbations, respectively. Then $\mathcal{K}_{res} \subseteq \mathcal{K}_{uni}$.*

*Proof.* Let $K_{\text{res}}(z_1, \ldots, z_n) = \sum_{i=1}^{n} z_i \cdot K_i + (1 - \sum_{i=1}^{n} z_i) \cdot Id = \sum_{i=1}^{n} z_i \cdot (K_i - Id) + Id$. We construct kernel parameters $Z^* = [z^*_{11}, z^*_{12}, \ldots, z^*_{k_1 k_2}]$ where $z^*_{pq} = \sum_{i=1}^{n} z_i \cdot (K_i - Id)_{pq}$. Then, $K_{\text{uni}}(Z^*) = \sum_{p,q} U_{pq} \cdot z^*_{pq} + Id = \sum_{i=1}^{n} z_i \cdot (K_i - Id) + Id = K_{\text{res}}$, showing any restricted kernel can be expressed as a universal kernel. □

**Thm. 1** (Perturbation Space Containment)**.** *Let $\mathcal{P}_{res} = \{X * K : K \in \mathcal{K}_{res}\}$ be the set of all perturbed images generated by restricted convolutional perturbation, and $\mathcal{P}_{uni} = \{X * K : K \in \mathcal{K}_{uni}\}$ be the set of all perturbed images generated by universal convolutional perturbation. Then $\mathcal{P}_{res} \subseteq \mathcal{P}_{uni}$.*

*Proof.* Since $\mathcal{K}_{\text{res}} \subseteq \mathcal{K}_{\text{uni}}$ (see Lem. 1), for any perturbed image $X * K$ where $K \in \mathcal{K}_{\text{res}}$, we have $K \in \mathcal{K}_{\text{uni}}$, which implies $X * K \in \mathcal{P}_{\text{uni}}$. Therefore, $\mathcal{P}_{\text{res}} \subseteq \mathcal{P}_{\text{uni}}$. □

Tab. 1. Discrete kernel results (UNSAT/SAT/TIMEOUT).

| | **Perturbation** | **Motion 0°** | **Motion 15°** | **Motion 30°** | **Motion 45°** | **Motion 60°** | **Motion 75°** | **Motion 90°** |
|---|---|---|---|---|---|---|---|---|
| VeriDou$_{\alpha\beta}$ | Restricted | 169/73/58 | 164/85/51 | 148/93/59 | 138/105/57 | 154/88/58 | 171/73/56 | 169/81/50 |
| | Universal | 135/120/45 | 134/117/49 | 127/126/47 | 122/128/50 | 134/125/41 | 135/117/48 | 125/132/43 |
| | Dual ($C = 0.5$) | 75/163/62 | 79/170/51 | 75/170/55 | 65/178/57 | 73/173/54 | 81/172/47 | 86/172/42 |
| | Dual ($C = 1.0$) | 50/206/44 | 51/203/46 | 51/211/38 | 48/219/33 | 51/208/41 | 49/201/50 | 49/198/53 |
| VeriDou$_{NS}$ | Restricted | 183/69/48 | 176/82/42 | 170/91/39 | 156/100/44 | 172/83/45 | 193/73/34 | 185/80/35 |
| | Universal | 126/120/54 | 123/117/60 | 116/125/59 | 113/129/58 | 121/125/54 | 121/117/62 | 114/132/54 |
| | Dual ($C = 0.5$) | 81/162/57 | 83/169/48 | 79/170/51 | 71/179/50 | 89/170/41 | 91/169/40 | 88/171/41 |
| | Dual ($C = 1.0$) | 56/203/41 | 58/199/43 | 57/206/37 | 53/215/32 | 59/205/36 | 59/199/42 | 56/195/49 |
| VeriDou$_{VS}$ | Restricted | 142/34/124 | 136/40/124 | 129/40/131 | 121/46/133 | 138/35/127 | 151/23/126 | 142/35/123 |
| | Universal | 113/116/71 | 112/115/73 | 108/122/70 | 102/125/73 | 116/124/60 | 115/117/68 | 111/131/58 |
| | Dual ($C = 0.5$) | 60/150/90 | 60/160/80 | 60/161/79 | 51/167/82 | 59/165/76 | 68/162/70 | 66/164/70 |
| | Dual ($C = 1.0$) | 28/192/80 | 26/194/80 | 24/198/78 | 24/203/73 | 28/195/77 | 29/190/81 | 33/186/81 |

Tab. 2. Continuous kernel results (UNSAT/SAT/TIMEOUT).

| | **Perturbation** | **0 → 30°** | **30 → 60°** | **60 → 90°** |
|---|---|---|---|---|
| VeriDou$_{\alpha\beta}$ | Restricted | 189/48/33 | 172/62/36 | 187/53/30 |
| | Universal | 98/136/36 | 88/115/67 | 103/129/38 |
| | Dual ($C = 0.5$) | 71/153/46 | 63/132/75 | 84/142/44 |
| | Dual ($C = 1.0$) | 51/168/51 | 47/138/85 | 54/171/45 |
| VeriDou$_{NS}$ | Restricted | 190/47/33 | 186/58/26 | 191/52/27 |
| | Universal | 91/127/52 | 89/111/70 | 101/119/50 |
| | Dual ($C = 0.5$) | 74/139/57 | 66/127/77 | 85/136/49 |
| | Dual ($C = 1.0$) | 56/155/59 | 53/135/82 | 59/161/50 |
| VeriDou$_{VS}$ | Restricted | 156/20/94 | 154/23/93 | 161/19/90 |
| | Universal | 87/121/62 | 76/101/93 | 97/112/61 |
| | Dual ($C = 0.5$) | 54/128/88 | 44/107/119 | 67/122/81 |
| | Dual ($C = 1.0$) | 33/134/103 | 28/113/129 | 37/136/97 |

Tab. 3. Arbitrary kernel results (UNSAT/SAT/TIMEOUT).

| | **Num. $z_i$ (%)** | **Domain 1** | **Domain 2** | **Domain 3** |
|---|---|---|---|---|
| VeriDou$_{\alpha\beta}$ | 20 | 132/81/87 | 121/92/87 | 114/99/87 |
| | 50 | 99/73/128 | 91/73/136 | 99/67/134 |
| | 100 | 74/58/168 | 81/50/169 | 73/65/162 |
| VeriDou$_{NS}$ | 20 | 143/81/76 | 135/91/74 | 126/99/75 |
| | 50 | 99/73/128 | 94/73/133 | 102/67/131 |
| | 100 | 72/58/170 | 78/51/171 | 70/65/165 |
| VeriDou$_{VS}$ | 20 | 117/81/102 | 105/92/103 | 101/99/100 |
| | 50 | 93/73/134 | 81/72/147 | 94/68/138 |
| | 100 | 68/57/175 | 78/51/171 | 71/65/164 |